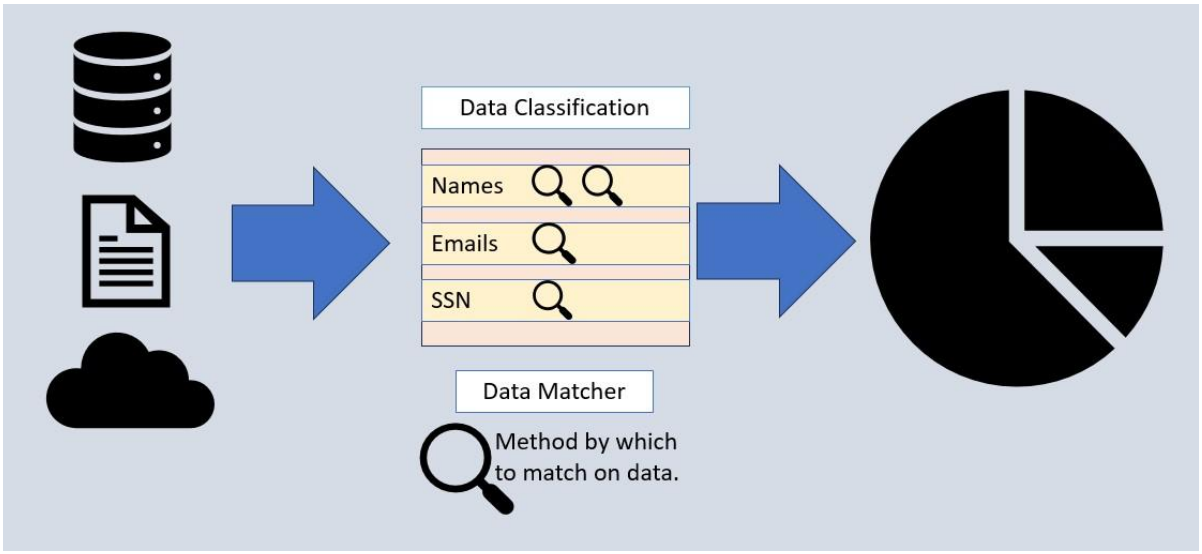
New IRI Data Classification Infrastructure

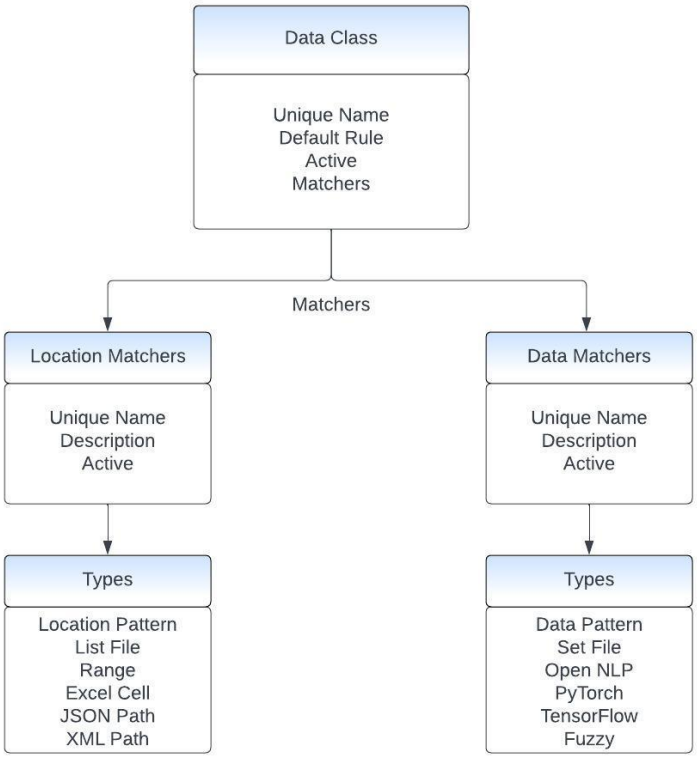


Data classification unifies the definition, discovery, and treatment of specific kinds, or classes, of data – usually PII like email addresses, last names, address, credit card or ID numbers, and so on. For multiple applications in the IRI Voracity platform – including most FieldShield and DarkShield data masking jobs – data classes describe, locate, and apply common functions (as rules) to those data classes, which often exist in multiple sources and silos.

As of May 2023, IRI data classes are now stored in a file called *iriLibrary.dclib* and are found within a project folder. This allows you to create either a “master” data class library that can be used for different tasks, or to create a library specific to a job.

The structure of a data class is:

- Unique name
- Description (optional)
- Default Rule (optional)
- Active
- Matchers
    - Location
    - Data



Unique Name and Description

Data class names need to start with a character. Dashes and underscores are allowed. Empty spaces and any other special characters are not allowed. You can add a description for any needed context for what the data class is used for, again this is optional.

Default Rule

A default rule is useful to ensure that a specific rule will be used with a data class, regardless of what task or job you are performing.

There can be situations where the default rule is not appropriate for a task, which is why you have granular control to override the default rule and select a different rule for a specific source.

Active

The *Active* attribute is used to filter data classes when they populate in a wizard as an option when classifying a source. If you do not want to use a data class for future jobs but do not want to delete it, deselect the active attribute.

Matchers

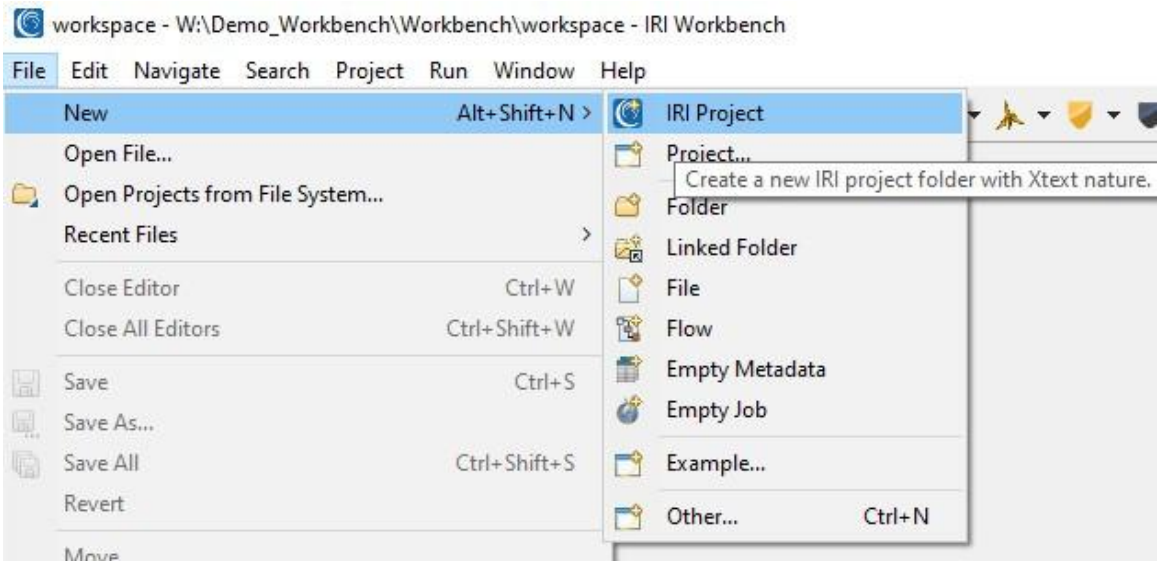
In the various data discovery operations supported in Voracity, data classes can be identified (PII can be found) using two different types of search matchers. Data matchers look at the data to find and verify the PII within each source, while location matchers look at the structure of the source itself to find PII. Each data class needs at least one matcher for auto-classification.

Matchers also have an active attribute which allows you to control which matchers are used during classification without the need to delete them.

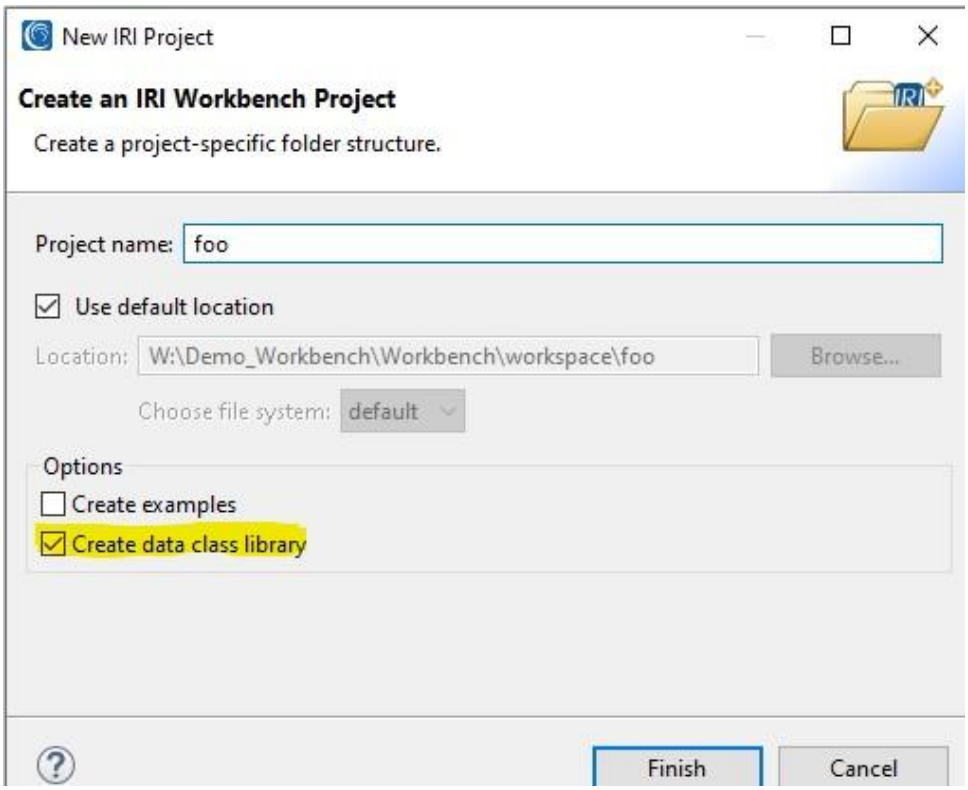
See Search Matchers below for more information.

Creating Data Classes

Data classes are available in the data class library. One is created by default when you create a new project folder in IRI Workbench. To create a new IRI Project, at the top of Workbench select *File > New > IRI Project*.

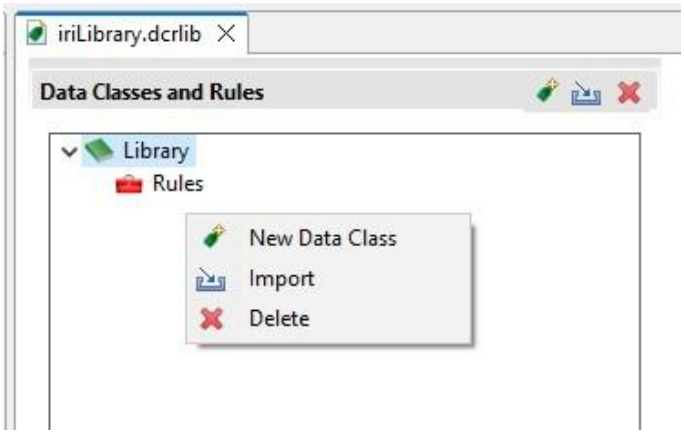


Give a unique name to your project. In the option section below, the *create data class library* is selected by default. If a project does not need a data class library, deselect this option.



To start creating data classes and rules, double-click on the *iriLibrary.dcrlib* file in the project folder that was created. This will open the editor on the right side of the Project Explorer and allow you to create, edit, and remove data classes and rules as well.

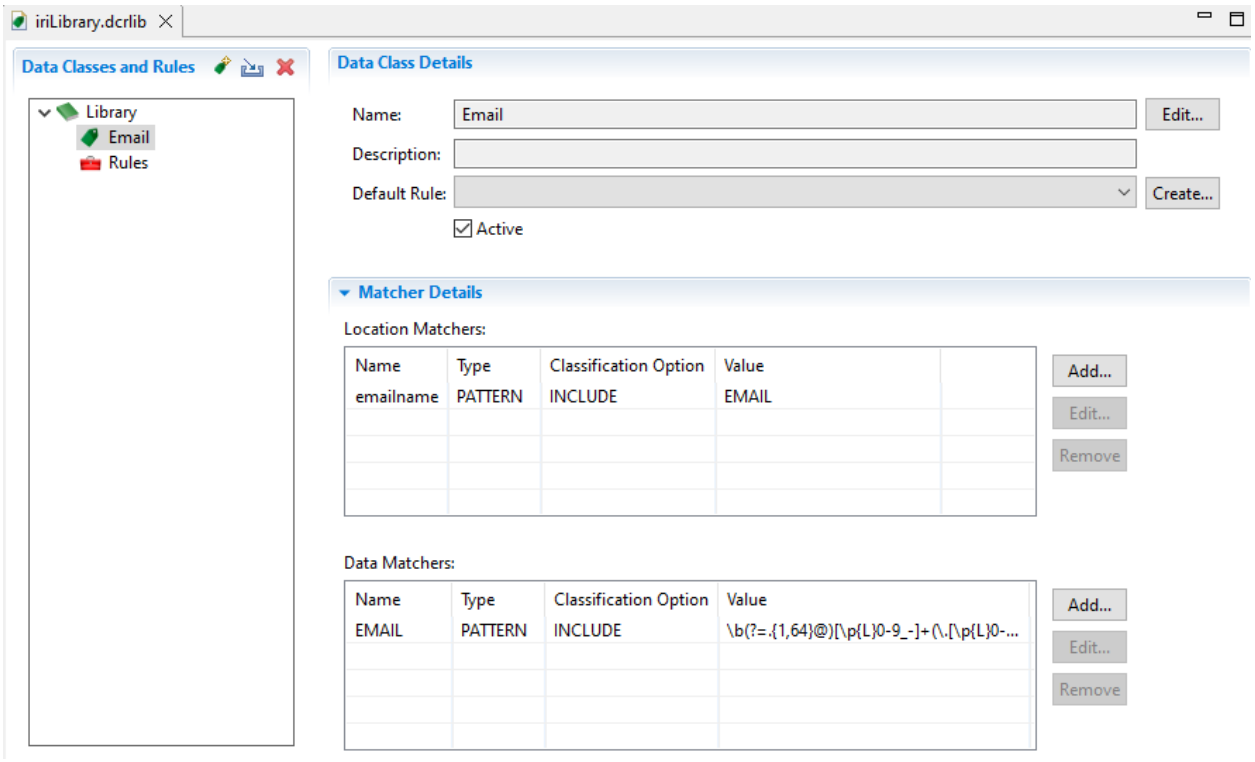
To create a data class, select the green button *(new data class)* at the top next to the title Data Classes and Rules or right click inside the outline and select the *New Data Class* option. This will open up a dialog allowing you to give a unique name to the data class.



Data class names need to start with a character. Dashes and underscores are allowed. Empty spaces and any other special characters are not allowed. You can add to the description section for any needed context for what the data class is used for; again however, this is optional. Select OK to create the data class.

The new data class will appear on the left side underneath the *Library*. Click on the data class and the right side of the form editor (called the data class details page) provides you with the ability to configure the data class.

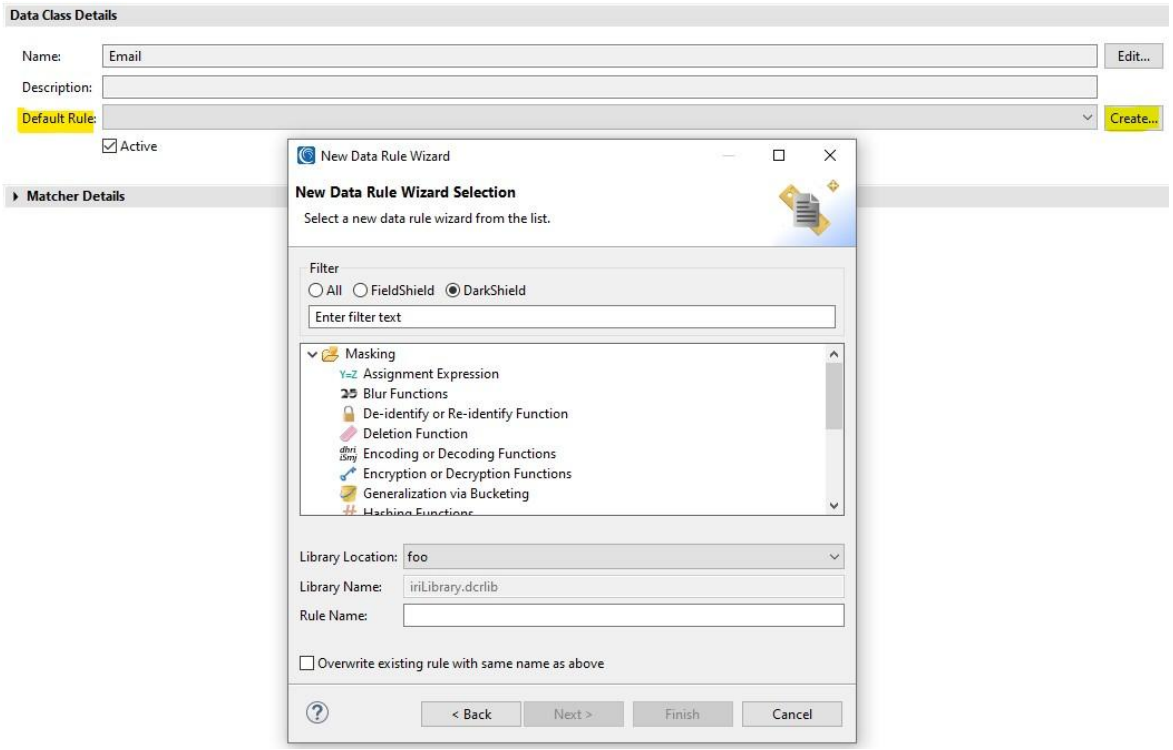
In the example below a data class called email was created.



Inside the data class details page, you can change the name of the data class, add a default rule, and create matchers to find PII in your data source.

Default Rule

To create a default rule, select the *Create…* button to the right of the Default Rule label and a dialog will appear. There are three major types of rules: Data, Quality, and Section. Data rules is the most common type. At the top of the dialog there is a filter section to expose rules which only relate to DarkShield.



Once a rule is created, select the drop down menu next to the Default Rule label and select a rule to be the default (masking function) for that data class.

Search Matchers

Both DarkShield and FieldShield use Matchers, or Search Matchers, to find PII (or any other desired data) in your data source(s) and create a map between the data and the data class to then apply your rule for masking that data. Search matchers are either *Data* or *Location* matchers. Location matchers are faster compared to data matchers, but are limited to when/how a data source is structured.

Example: A location matcher for emails would look for emails in a DB, CSV or Excel column with “email” in the name of the column. But if a column called “contact” contains email addresses, the “structure” is different from what the matcher expects, so you will not get a match on that.

Data matchers are slower but find PII through data (format/content) directly, and can be used with multiple sources without knowing if/how it is structured. From the above example, columns can have any name and not prevent a data matcher from finding email addresses and creating a match because the data matcher would evaluate the data itself with a RegEx pattern for email.

By using both location and data matchers simultaneously, you can ensure that PII will be found in your data source by either the structure or data format/s it contains. Note as well that you can also use more than one location matcher and data matcher at the same time for even more certainty, but the more matching attempts you specify, the longer data discovery can take.

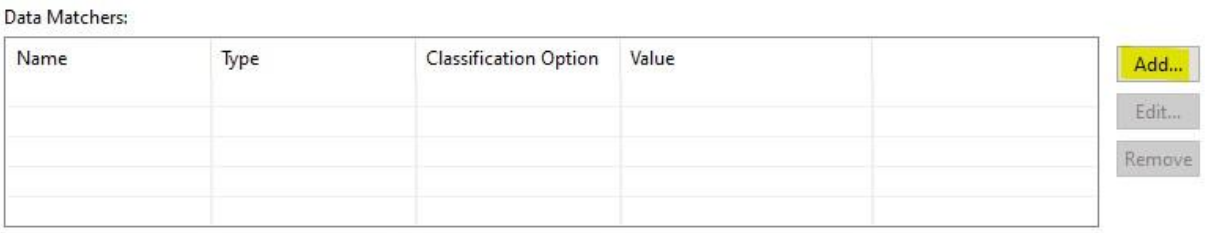
Data Matchers

Data matchers, as the name suggests, look at the actual data format (content) or grammatical context to verify if it belongs to your data class (is PII). Currently there are six different types of data matchers:

1. Data Pattern (regular expressions)
2. Set File (value lookup)
3. Open NLP (NER, DarkShield only)
4. PyTorch (NER, DarkShield only)
5. TensorFlow (NER, DarkShield only)
6. Fuzzy (match to values in a set file, DarkShield only)

To add a data matcher to the data class, select the add button and a dialog will appear, allowing you to create different types of matchers.

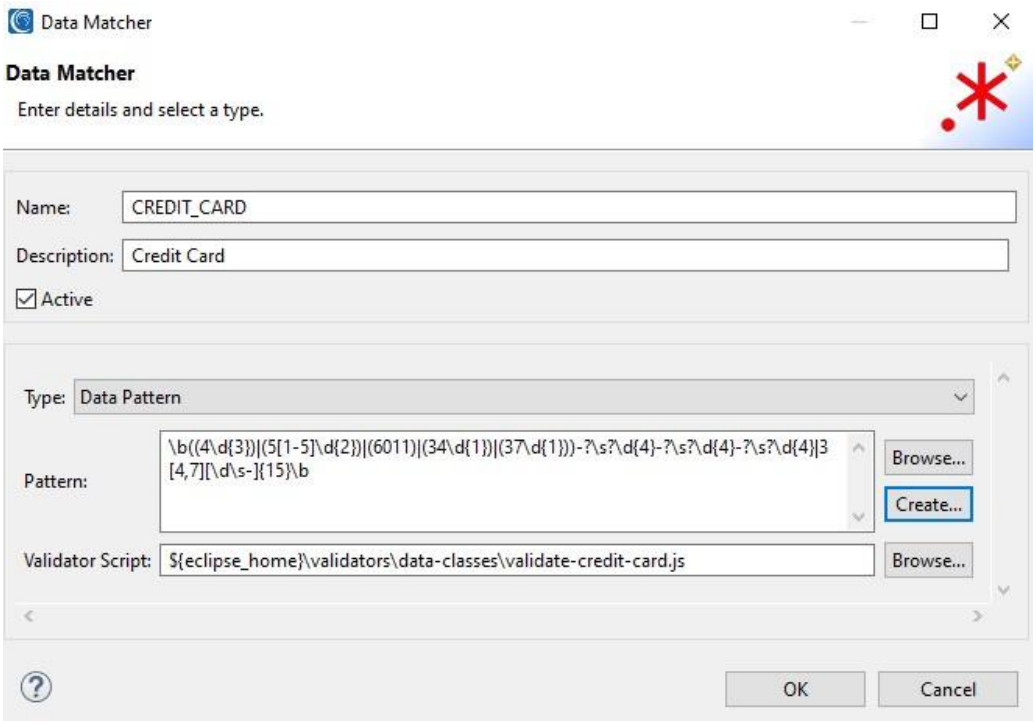
You are able to create multiple matchers as needed but there are some pros and cons that you should consider. The more data matchers that are created for a single data class can ensure that specific PII is found. The con however, the more matchers you have the more computation is done to find PII which can affect the time classification takes.



All data matchers need a unique name, can have a description, and have an active attribute that works the same as the data class active attribute.

To select a different type of data matcher, select the drop down menu next to the type label. Depending on the type of matcher, attributes required for the specific matcher will appear below.

Data patterns (a type of data matcher) can be easily created. Workbench comes with a library of regular expressions and validator scripts already available. The image below shows a credit card pattern and validator script that is included with Workbench.



Location Matchers

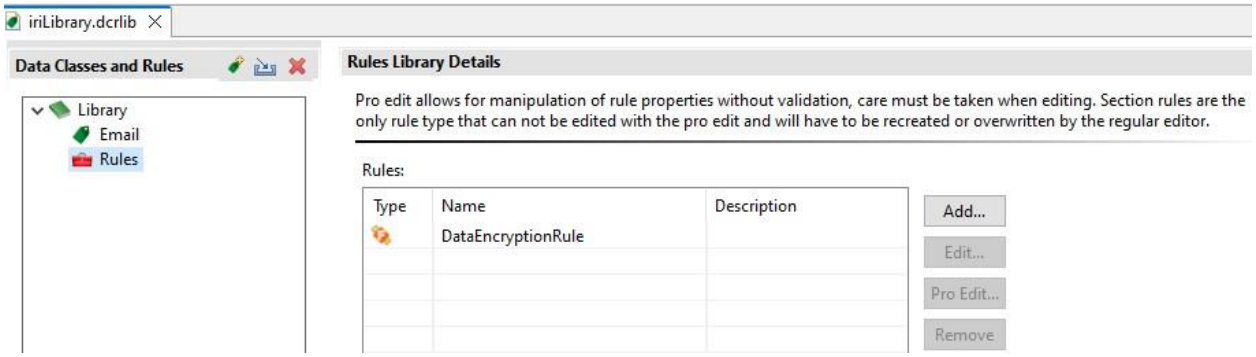
The second type of matcher is a location matcher. These matchers look at the structure of the *source* (but not the data itself) in order to find PII. There are currently 6 types:

1. Location Pattern (regular expression for the a column name, for example)
2. List File (matches a value in a list of locations, like a list of column names)
3. Range (matches on a range of indexed locations; see details in the Appendix)
4. Excel Cell (DarkShield Only)
5. JSON Path (DarkShield Only)
6. XML Path (DarkShield Only)

Creating, editing, and deleting location matchers is the same process as the data matchers.

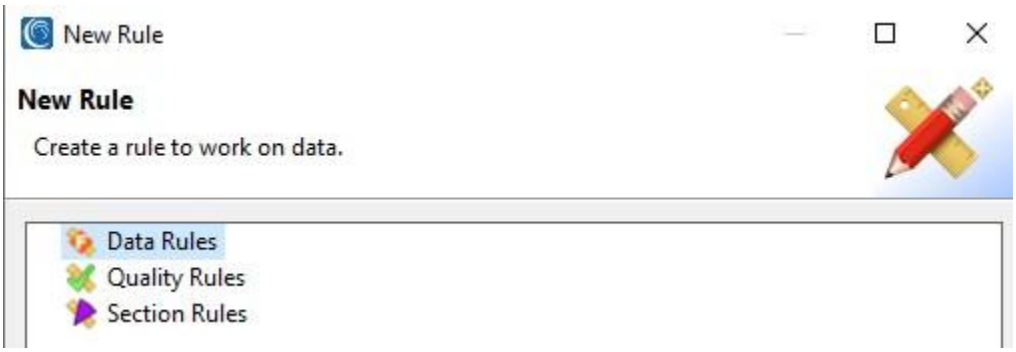
Rule Library (for Data Masking functions)

Assigning a consistent masking function to your data classes (as a rule) preserves data and referential integrity in the target(s), making test data more reliable, and in some cases, more reversible. Since data classes can have default rules, the rule library is stored within the data class library to prevent the deletion of rules (previously stored separately) that are referenced by a data class. You should use this new rule library to create, edit and delete your (masking) rules.

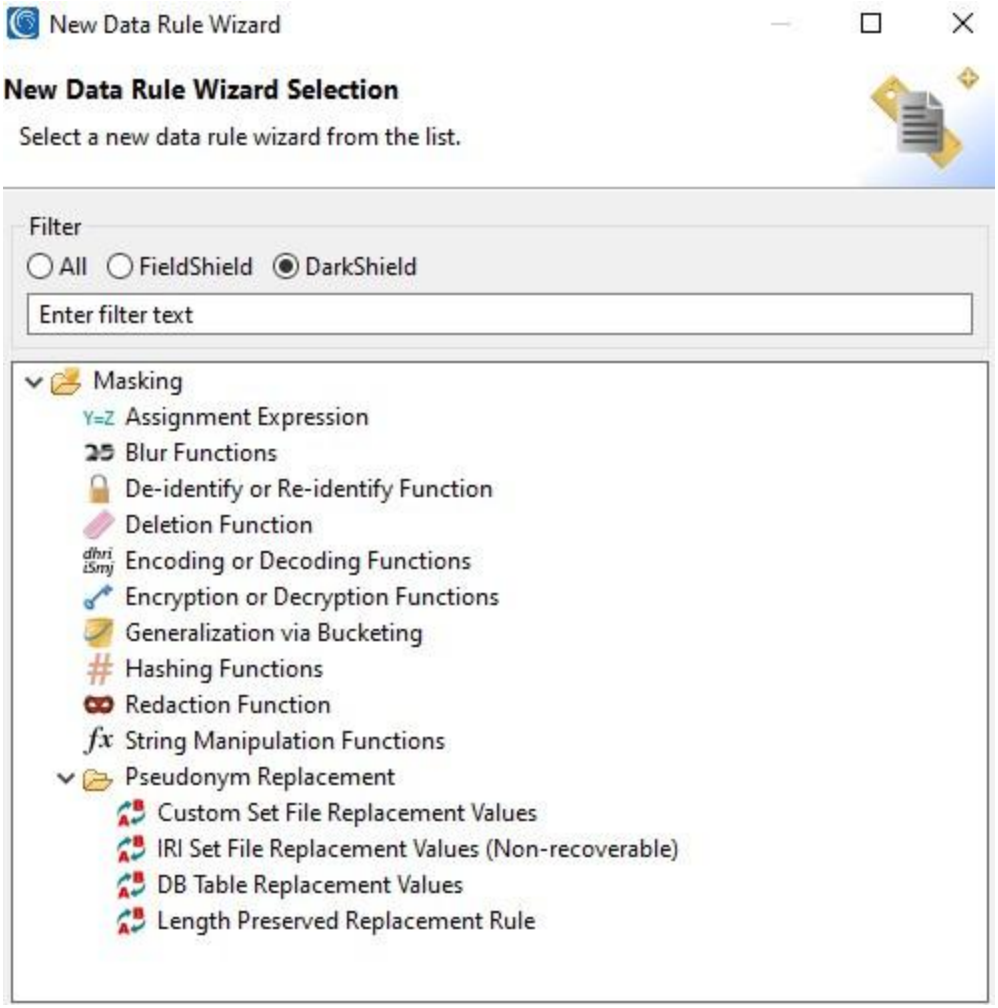


Add

To add rules to the library click on the *Add…* button and a dialog will appear with three different rule types to choose from which are data, section and quality rules. Data rules are the most common one you will use. Select the rule type you want to create and select next at the bottom of the dialog to create a rule.

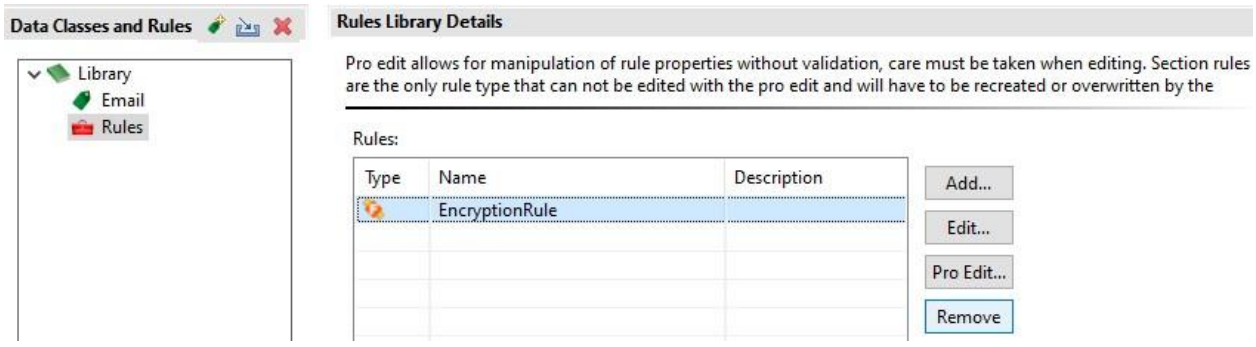


When selecting the data rule you can filter which rules are available to you depending on the product you will be using. Select the *DarkShield* radio button to filter rules that are available for you to use.



Remove

In order to remove a rule from the rules library, select the red tool box called *Rules* in the editor and select the rule you wish to remove. The selected rule will be highlighted and the buttons to the right will be enabled. Click on remove and the rule will be deleted from the library.



When you delete a rule, any data class that uses that rule as a default will no longer have a default rule associated with the data class.

Edit Rules

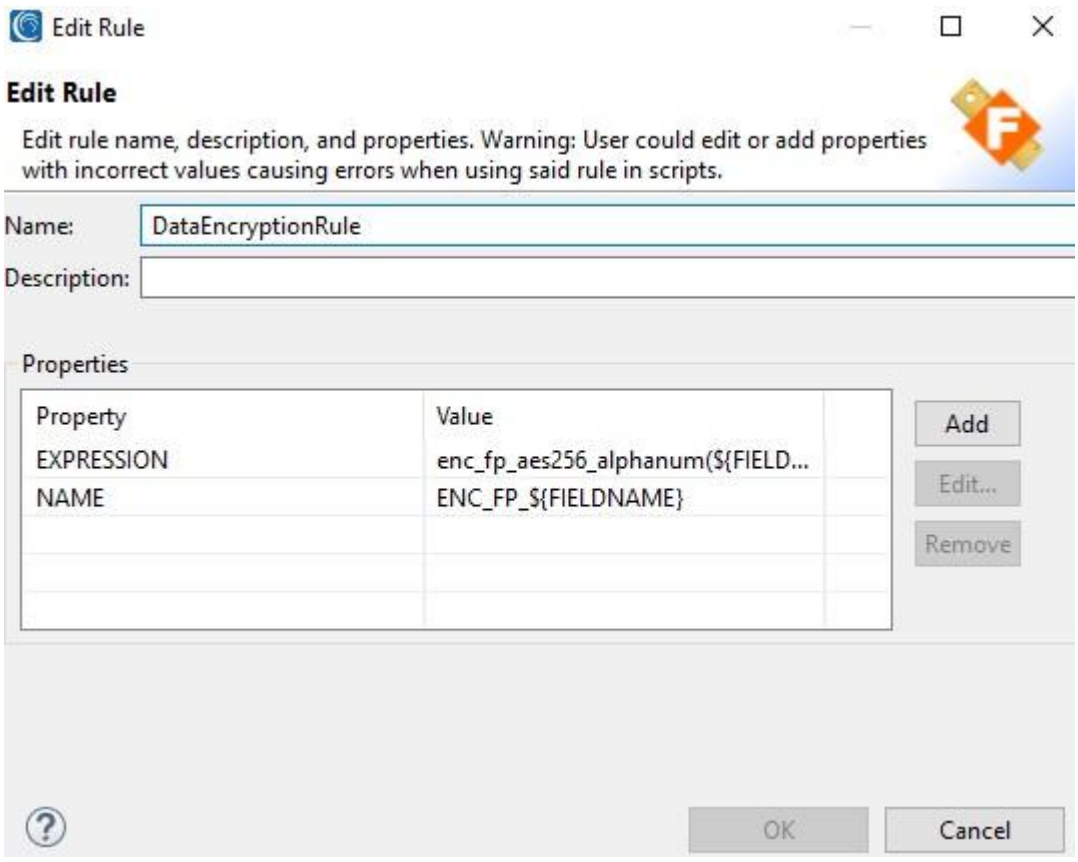
Select a rule to enable the option to edit or remove the selected rule. There are two different options when it comes to editing.

The *Edit…* option opens the wizard that created the rule to allow you to modify its properties and overwrite the rule. If there is an issue with opening the wizard that originally created the rule, then the Pro Edit dialog will appear.

Note: Due to how *section* rules are created and saved, they cannot be used with the Pro Edit wizard. Also, the regular Edit wizard can only overwrite the rule. It will not populate properties that were used to create the rule originally.

Pro Edit

The *Pro Edit …* option allows you to edit and add properties that are not exposed using the regular wizards. Since there is no hand-holding or validation of what is specified, once the changes are saved, the rule can only be edited with Pro Edit going forward.



[1] ${FIELDNAME} in the rule definition is simply a placeholder. In the context of DarkShield, ${FIELDNAME} is a piece of data that was matched.